



Weighing the “Heaviest” Polya Urn

Jeremy Chen

*Department of Decision Sciences, National University of Singapore Business School
15 Kent Ridge Drive, Singapore 119245*

Abstract

For the classical Polya urn model parametrized by exponent γ , the limit distributions of the fraction of balls in each bin are simple and well known when $\gamma < 1$ (“egalitarian”) and when $\gamma > 1$ (“winner takes all”). In this note, we partially fill in the gap for $\gamma = 1$, the critical point, by providing explicit analytical expressions for all the moments of the limit distribution of the fraction of balls in the bin with the most balls.

© Jeremy Chen 2014; Last Updated: June 25, 2014

Keywords: Polya urn, Proportional Preferential Attachment, Limit distribution, Herd Behaviour

1. Preliminaries

The Polya urn problem describes a well-studied family of random processes that have been fruitfully applied in diverse fields ranging from telecommunications to understanding self-organizing processes like network formation and herd behavior. In the classical Polya urn problem, one begins with d bins, each containing one ball. Additional balls arrive one at a time, and the probability that an arriving ball is placed in a given bin is proportional to m^γ , where m is the number of balls in that bin.

In this note, we consider the case of $\gamma = 1$, which corresponds to a process of “proportional preferential attachment” and is a critical point with respect to the limit distribution of the fraction of balls in each bin. It is well known that for $\gamma < 1$ the fraction of balls in the “heaviest” bin (the bin with the most balls) tends to $1/d$, and for $\gamma > 1$ the fraction of balls in the “heaviest” bin tends to 1. (See, for instance, surveys such as [1] and [2] or books such as [3] and [4].) Unexpectedly, though scientists and engineers are interested in analogous quantities such as “the size of the largest (biological) plague”, this question has not been explored for the case of $\gamma = 1$.

The modest contribution of this note is to explore that very problem. We characterize the limit distribution of the fraction of balls in the “heaviest” bin for $\gamma = 1$ by providing explicit analytical expressions for all its moments.

2. The Main Result

Denote the number of balls in the “heaviest” urn (when there are d urns), after a total of $n - d \geq 0$ balls are added, as $H_d(n)$. For integer $m \geq 0$ and integer $d \geq 1$, let

$$M_d^{(m)} := \lim_{n \rightarrow \infty} \mathbb{E} \left[\left(\frac{H_d(n)}{n} \right)^m \right]. \quad (1)$$

Email address: jeremy.chen@nus.edu.sg (Jeremy Chen)

16 For notational convenience, let $M_0^0 := 1$. We note that there are three equivalent ways of describing $M_d^{(m)}$:

Proposition 2.1 (Limiting Moments of the Fraction of Balls in the “Heaviest” Urn). For $m \geq 1$ and $d \geq 2$,

$$M_d^{(m)} = \sum_{k=0}^m \frac{d-1}{d^k} \frac{m!}{(m-k)!} \frac{(m+d-k-2)!}{(m+d-1)!} M_{d-1}^{(m-k)}, \quad (2)$$

$$M_d^{(m)} = \frac{d-1}{m+d-1} M_{d-1}^{(m)} + \frac{m}{d(m+d-1)} M_d^{(m-1)}, \text{ and} \quad (3)$$

$$M_d^{(m)} = \frac{m}{m+d-1} \sum_{j=1}^d \frac{(d-1)!}{j!} \frac{(m+j-2)!}{(m+d-2)!} M_j^{(m-1)}. \quad (4)$$

17 This result is an explicit characterization in the sense that the computations to be done using either of equations
18 (2) thru (4) are explicit (typically simple arithmetic) rather than implicit (e.g.: solving a system of equations). One
19 readily observes that this is so because the next $M_d^{(m)}$ to be computed depends only on already computed moments. (It
20 is clear that $M_d^{(0)} = 1$ for all $d \geq 1$ and $M_1^{(m)} = 1$ for all $m \geq 1$.)

21 The first and second moments being typically of special interest, we may use equation (4) to compute them:

Corollary 2.2 (Limiting Mean/Second Moment of the Fraction of Balls in the “Heaviest” Urn).

$$\frac{1}{d} \log d \leq M_d^{(1)} = \frac{1}{d} \sum_{k=1}^d \frac{1}{k} \leq \frac{1}{d} (\log d + 1) \quad (5)$$

and

$$M_d^{(2)} = \frac{2}{d(d+1)} \sum_{j=1}^d M_j^{(1)} = \frac{2}{d(d+1)} \sum_{j=1}^d \frac{1}{j} \sum_{k=1}^j \frac{1}{k}. \quad (6)$$

22 Both “scale like” $d^{-1} \log d$. Direct computation reveals that the coefficient of variation (the ratio of the standard
23 deviation to the mean) of the fraction of balls in the “heaviest” urn rises to a peak of about 0.27 at $d = 10$ after which
24 it begins to decline with increasing d to a limiting value of 0.

25 More generally, though the recurrences enable us to obtain expressions for any given moment by iterated substi-
26 tution, there does not appear to be any special structure that enables drastic simplification. This is unfortunate.

27 Before proceeding to the prove Proposition 2.1, we reproduce, for completeness, the following well known result
28 on the distribution of the number of balls in each bin:

Lemma 2.3. The number of balls in the d urns after m additional balls are added is uniformly distributed on

$$S_{(d,m)} := \{v \in \mathbb{Z}^d : v \geq e, v^T e = d + m\}. \quad (7)$$

Furthermore,

$$|S_{(d,m)}| = \frac{(m+d-1)!}{m!(d-1)!} = \binom{m+d-1}{d-1}. \quad (8)$$

Proof of Lemma 2.3: Let the probability that v_k balls are in the k -th urn (of d) after m additional balls are added be $\pi_{(d,m)}(v)$. Clearly, $\pi_{(d,0)}(e) = 1$. If $\pi_{(d,m)}(v) = \eta_{(d,m)}$ for some constant $\eta_{(d,m)}$ for all $v \in S_{(d,m)}$. Then, for all $v \in S_{(d,m+1)}$,

$$\begin{aligned} \pi_{(d,m+1)}(v) &= \sum_{v_k > 1} \eta_{(d,m)} \frac{v_k - 1}{m + d} \\ &= \sum_{k=1}^d \eta_{(d,m)} \frac{v_k - 1}{m + d} \\ &= \eta_{(d,m)} \frac{m + 1}{m + d} \\ &=: \eta_{(d,m+1)} \end{aligned} \quad (9)$$

where the first equality follows from the dynamics of preferential attachment. Now, $|S_{(d,m)}| = 1/\eta_{(d,m)}$, and $|S_{(d,0)}| = 1$. Therefore, by equation (9),

$$S_{(d,m)} = 1 \cdot \frac{d}{1} \cdot \frac{d+1}{2} \cdots \frac{m+d-1}{m}$$

and the proof is complete. \blacksquare

Subsequently, Lemma 2.3 and a simple partitioning of the set of possible outcomes ($S_{(d,m)}$, a discrete simplex) will be used to characterize the limiting distribution of the fraction of balls in the “heaviest” urn. We will also make use of the easily verifiable fact that:

Lemma 2.4. For integer $a, b > 0$ and real-valued $c > 0$,

$$\int_0^c x^a (c-x)^b dx = \frac{a!b!}{(a+b+1)!} c^{a+b+1}. \quad (10)$$

Proof of Proposition 2.1: Clearly, $M_d^{(0)} = 1$ and $M_1^{(m)} = 1$. For cases where $m > 0$ and/or $d > 1$, an expression for the desired moment for finite n will be constructed, and the limit as $n \rightarrow \infty$ evaluated.

Now, the set $S_{(d,(\alpha-1)d)}$ (for $\alpha \in \mathbb{N}$), as defined in Lemma 2.3, may be expressed as the following disjoint union:

$$S_{(d,(\alpha-1)d)} = \{\alpha e\} \cup \bigcup_{\mu=1}^{\alpha-1} \bigcup_{\tau=1}^{d-1} T_{(\alpha,d,\mu,\tau)}$$

where $T_{(\alpha,d,\mu,\tau)} := \{v \in \mathbb{Z}^d : v \geq \mu e, v^T e = \alpha d, \gamma(v, \mu) = \tau\}$ and $\gamma(v, x) := |\{k : v_k = x\}|$ is the number of entries of the vector v with the value x . This is so because there is a single vector in $S_{(d,(\alpha-1)d)}$ where all entries take the same value (αe), and $\alpha-1$ other possible values of the smallest entry of vectors in $S_{(d,(\alpha-1)d)}$ (specifically, $1, 2, \dots, \alpha-1$). In each of the latter cases, the number of entries taking on the minimum value may range from 1 to $d-1$.

Noting that vectors in $T_{(\alpha,d,\mu,\tau)}$ each have τ entries taking value μ , and $d-\tau$ entries taking values strictly larger than μ , the cardinality of $T_{(\alpha,d,\mu,\tau)}$, must be $|S_{(d-\tau,(\alpha-\mu)d-(d-\tau))}|$ multiplied by the number of ways to pick the τ entries taking value μ . Therefore, using Lemma 2.3, one may deduce that

$$|T_{(\alpha,d,\mu,\tau)}| = |S_{(d-\tau,(\alpha-\mu)d-(d-\tau))}| \binom{d}{\tau} = \binom{(\alpha-\mu)d-1}{d-\tau-1} \binom{d}{\tau}. \quad (11)$$

Thus one obtains the identity $\binom{\alpha d-1}{d-1} = 1 + \sum_{\mu=1}^{\alpha-1} \sum_{\tau=1}^{d-1} \binom{(\alpha-\mu)d-1}{d-\tau-1} \binom{d}{\tau}$ which, itself, may be proven directly by induction.

Now, equation (1) may be written equivalently as

$$\mathbb{E}[H_d(n)^m] = M_d^{(m)} n^m + o(n^m) \quad (12)$$

for integer $m \geq 1$. Equation (12) clearly holds for $d = 1$. Now, suppose that equation (12) is true for $1, 2, \dots, d-1$ urns. Using the fact that conditional on realizations being in $T_{(\alpha,d,\mu,\tau)}$, the uniform probability of outcomes implies that the distribution of $H_d(n)$ is identical to the (unconditional) distribution of $H_{d-\tau}((\alpha-\mu)d) + \mu$, the m -th moment of the number of balls in the “heaviest” urn is given by

$$\begin{aligned} \mathbb{E}[H_d(n)^m | T_{(\alpha,d,\mu,\tau)}] &= \mathbb{E}[(H_{d-\tau}((\alpha-\mu)d) + \mu)^m] \\ &= \sum_{k=0}^m \binom{m}{k} M_{d-\tau}^{(m-k)} ((\alpha-\mu)d)^{m-k} \mu^k + o(\alpha^m) \end{aligned} \quad (13)$$

following an application of equation (12). Through equation (11), we obtain

$$\mathbb{E}\left[\left(\frac{H_d(\alpha d)}{\alpha d}\right)^m\right] = \frac{1}{(\alpha d)^m} \frac{1}{\binom{\alpha d-1}{d-1}} \left[\alpha^m + \sum_{\mu=1}^{\alpha-1} \sum_{\tau=1}^{d-1} \mathbb{E}[H_d(n)^m | T_{(\alpha,d,\mu,\tau)}] \binom{(\alpha-\mu)d-1}{d-\tau-1} \binom{d}{\tau} \right]. \quad (14)$$

Subsequently, taking limits,

$$\begin{aligned}
\lim_{\alpha \rightarrow \infty} \mathbb{E} \left[\left(\frac{H_d(\alpha d)}{\alpha d} \right)^m \right] &= \lim_{\alpha \rightarrow \infty} \frac{1}{(\alpha d)^m} \frac{1}{\binom{\alpha d - 1}{d - 1}} \left[\alpha^m + \sum_{\mu=1}^{\alpha-1} \sum_{\tau=1}^{d-1} \mathbb{E} [H_d(n)^m | T_{(\alpha, d, \mu, \tau)}] \binom{(\alpha - \mu)d - 1}{d - \tau - 1} \binom{d}{\tau} \right] \\
&= \lim_{\alpha \rightarrow \infty} \frac{\alpha^m + \sum_{\mu=1}^{\alpha-1} \sum_{\tau=1}^{d-1} \left[\sum_{k=0}^m \binom{m}{k} M_{d-\tau}^{(m-k)} ((\alpha - \mu)d)^{m-k} \mu^k + o(\alpha^m) \right] \left[\frac{((\alpha - \mu)d)^{d-\tau-1}}{(d - \tau - 1)!} + o(\alpha^{d-\tau-1}) \right] \binom{d}{\tau}}{(\alpha d)^m \frac{(\alpha d)^{d-1}}{(d-1)!}} \\
&= \lim_{\alpha \rightarrow \infty} \frac{\sum_{\mu=1}^{\alpha-1} \left(\left[\sum_{k=0}^m \binom{m}{k} M_{d-\tau}^{(m-k)} ((\alpha - \mu)d)^{m-k} \mu^k + o(\alpha^m) \right] \frac{(\alpha - \mu)^{d-2} d^{d-2}}{(d-2)!} d + o(\alpha^{m+d-2}) \right)}{(\alpha d)^m \frac{(\alpha d)^{d-1}}{(d-1)!}} \\
&= \lim_{\alpha \rightarrow \infty} \frac{d-1}{d^m \alpha^{m+d-1}} \sum_{\mu=1}^{\alpha-1} \left(\left[\sum_{k=0}^m \binom{m}{k} M_{d-1}^{(m-k)} ((\alpha - \mu)d)^{m-k} \mu^k + o(\alpha^m) \right] (\alpha - \mu)^{d-2} + o(\alpha^{m+d-2}) \right) \\
&= \lim_{\alpha \rightarrow \infty} \frac{d-1}{d^m \alpha^{m+d-1}} \sum_{\mu=1}^{\alpha-1} \left[\sum_{k=0}^m \binom{m}{k} M_{d-1}^{(m-k)} (\alpha - \mu)^{m+d-k-2} \mu^k d^{m-k} + o(\alpha^{m+d-2}) \right] \\
&= \lim_{\alpha \rightarrow \infty} \frac{d-1}{d^m \alpha^{m+d-1}} \left[\int_0^\alpha \sum_{k=0}^m \binom{m}{k} M_{d-1}^{(m-k)} (\alpha - \mu)^{m+d-k-2} \mu^k d^{m-k} + o(\alpha^{m+d-2}) d\mu + o(\alpha^{m+d-1}) \right] \\
&= \lim_{\alpha \rightarrow \infty} \frac{d-1}{d^m \alpha^{m+d-1}} \left[\sum_{k=0}^m \binom{m}{k} M_{d-1}^{(m-k)} \frac{(m+d-k-2)! k!}{(m+d-1)!} \alpha^{m+d-1} d^{m-k} + o(\alpha^{m+d-1}) \right] \\
&= \lim_{\alpha \rightarrow \infty} \left[\sum_{k=0}^m \frac{d-1}{d^k} \frac{m!}{(m-k)!} \frac{(m+d-k-2)!}{(m+d-1)!} M_{d-1}^{(m-k)} + o(1) \right] \\
&= \sum_{k=0}^m \frac{d-1}{d^k} \frac{m!}{(m-k)!} \frac{(m+d-k-2)!}{(m+d-1)!} M_{d-1}^{(m-k)}.
\end{aligned}$$

The seventh equality follows from an application of Lemma 2.4 to evaluate the integral.

To complete the proof, we argue that the limit

$$\lim_{r \rightarrow \infty} \mathbb{E} \left[\left(\frac{H_d(n_r)}{n_r} \right)^m \right] = \sum_{k=0}^m \frac{d-1}{d^k} \frac{m!}{(m-k)!} \frac{(m+d-k-2)!}{(m+d-1)!} M_{d-1}^{(m-k)}.$$

also arises for all increasing positive integer sequences $\{n_r\}_{r=1}^\infty$ whose elements are greater than d but are not all necessarily integral multiples of d .

Note that given any n , for $\beta_{x,d} := d \lfloor x/d \rfloor$, we have $\beta_{n,d} d \leq n \leq (\beta_{n,d} + 1)d$. Since, in the right-hand-side of equation (14), both the numerator and denominator are increasing in α , one can construct upper and lower bounds by using either $\beta_{n,d}$ or $\beta_{n,d} + 1$ accordingly in place of α in the numerator and denominator. For both bounds, the same leading order terms arise and hence the same limits. Therefore, with this sandwiching, we establish equation (2).

With equation (2), we may reduce the multiple term recurrence to a two term recurrence, yielding equation (3), and then perform iterative substitution to obtain equation (4). This proceeds as follows:

$$\begin{aligned}
 M_d^{(m)} &= \frac{d-1}{m+d-1} M_{d-1}^{(m)} + \frac{m}{d(m+d-1)} M_d^{(m-1)} \\
 &= \prod_{j=2}^d \frac{j-1}{m+j-1} + \sum_{j=2}^{d-1} \left(\prod_{k=j+1}^d \frac{k-1}{m+k-1} \right) \frac{m}{j(m+j-1)} M_j^{(m-1)} + \frac{m}{d(m+d-1)} M_d^{(m-1)} \\
 &= \frac{m!(d-1)!}{(m+d-1)!} + \sum_{j=2}^{d-1} m \frac{(d-1)!}{j!} \frac{(m+j-2)!}{(m+d-1)!} M_j^{(m-1)} + \frac{m}{d(m+d-1)} M_d^{(m-1)}.
 \end{aligned}$$

(The second equality holds because $M_1^{(m-1)} = 1$.) Simplifying yields equation (4), as desired. ■

3. Numerical Experiments

Illustrations of Proposition 2.1 may be found in Figures 1 thru 8 along with simulated average fractions of balls in the “heaviest” urn and quantiles obtained from simulation (for n ranging from 100 to 20,000). Those simulations suggest that the limiting values are good approximations even for systems with just thousands of balls added.

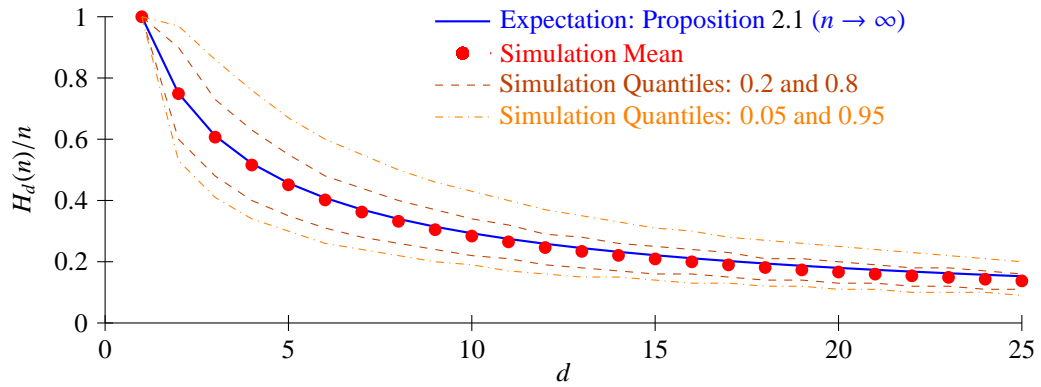
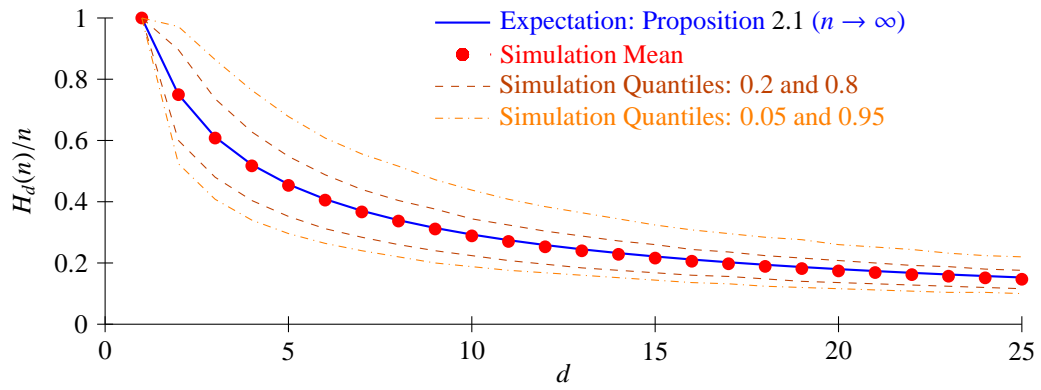
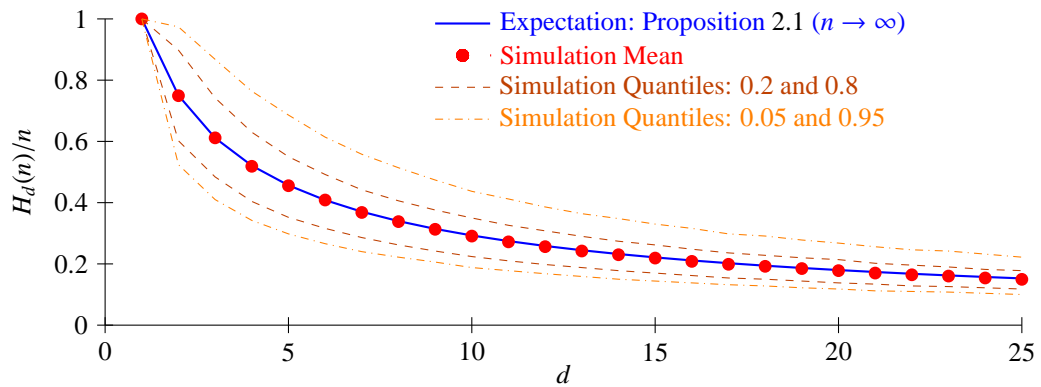
4. Herding by Preferential Attachment

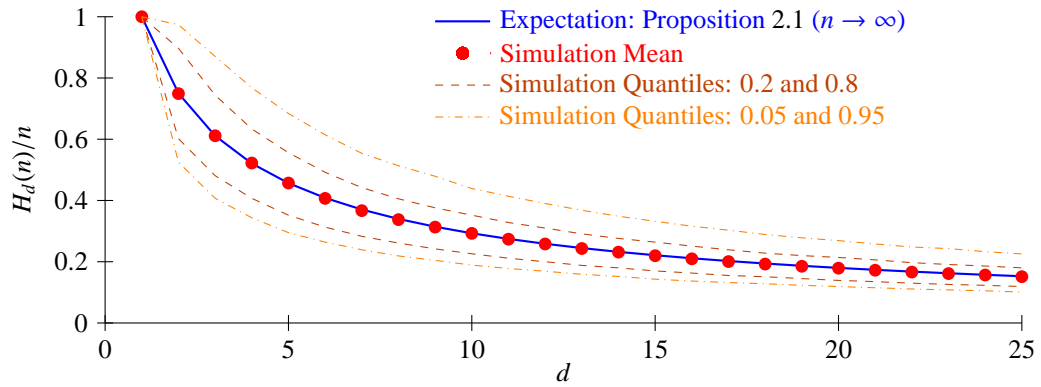
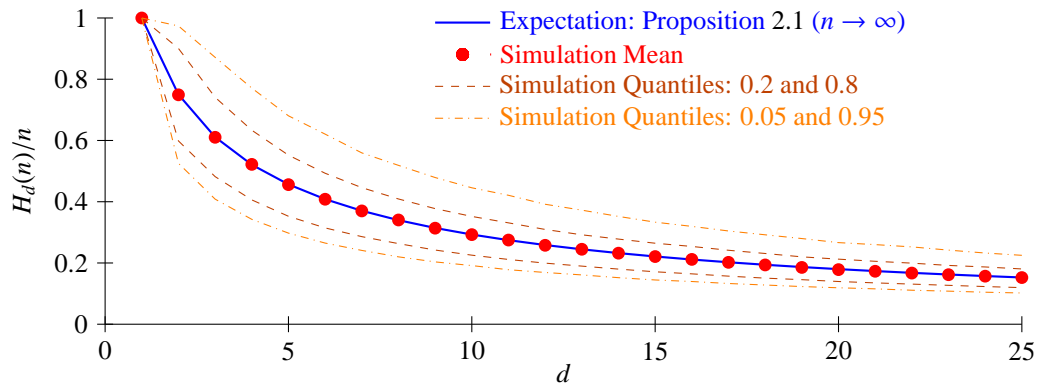
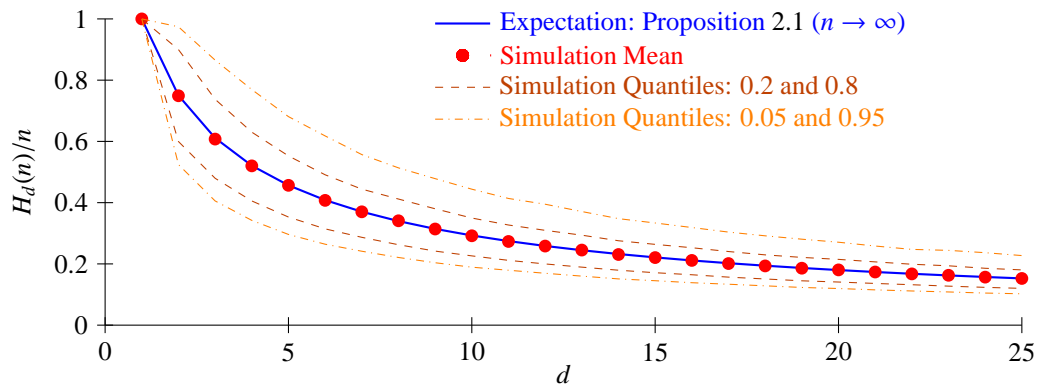
For the author, the particular question of what fraction of the balls, in an instance of a Polya urn problem, would end up in the urn with the most balls arose from studying a discrete choice model of connected agents whose choices affect those of others [5]. In that model, an agent would probabilistically select a choice based on his/her own preferences or adopt the choice of some other agent (i.e.: in “steady state”, the former agent makes choices in accordance with the choice probabilities of the other agent). By imposing specific decision making dynamics on that “steady state model”, a stochastic model was obtained whereby, in any sample path, the realized choice of any given agent might be traced “up the tree of choice adoption” to some “decisive agent” (an agent making his/her choices based on his/her own preferences rather than by adopting that of another agent). As such, each outcome would correspond to a “forest” of “trees” each containing exactly one “decisive agent” and rooted at that agent. Let us denote all the agents connected by choice adoption decisions to each “decisive agent” to be the “herd” associated with that “decisive agent”.

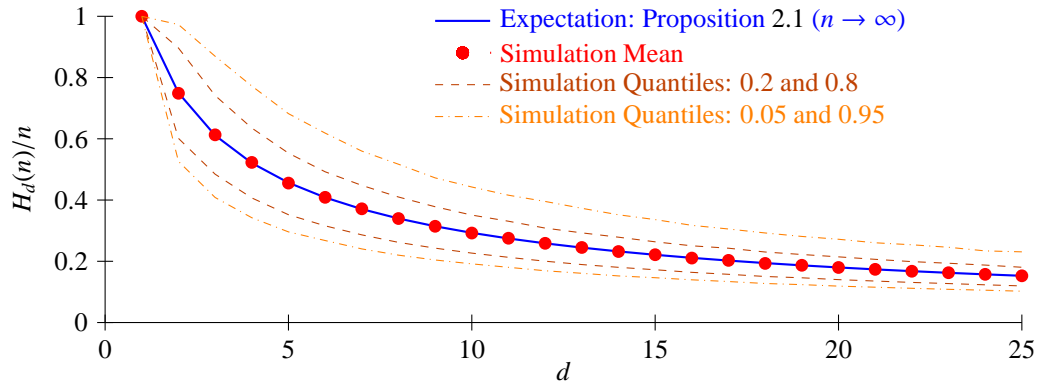
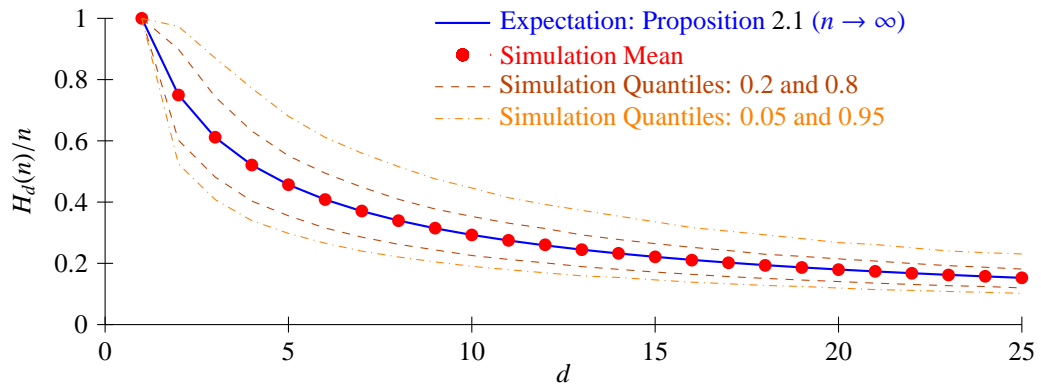
To gain some insight on herd behavior, the model might be simplified to one where, conditional on the “decisive agents”, any agent adopting the choice of another would select uniformly among the other agents. Further conditioning on there being d “decisive agents”, one obtains an instance of the Polya urn model with d urns. Using the largest herd as a proxy for the extent of herding behavior, Proposition 2.1 (and Corollary 2.2, in particular) inform one about the relationship between herding and the “decisiveness” of agents. Furthermore, should one desire to evaluate the “risk” associated with herding as the expectation of some (smooth) function of the fraction of agents in the largest herd, given that the moments are decreasing and bounded between 0 and 1 and that numerical experiments indicate the limiting distributions are “close” for reasonably sized populations of agents, good Taylor-expansion-based approximations can be computed.

References

- [1] F. Chung, S. Handjani, D. Jungreis, Generalizations of Polya’s urn Problem, *Annals of Combinatorics* 7 (2) (2003) 141–153.
- [2] R. Pemantle, A survey of random processes with reinforcement, *Probability Surveys* 4 (2007) 1–79.
- [3] N. L. Johnson, S. Kotz, *Urn models and their application: An approach to modern discrete probability theory*, Wiley, 1977.
- [4] H. M. Mahmoud, *Polya Urn Models*, CRC, 2008.
- [5] J. Chen, *Social Networks and the Choices People Make*, Working Paper (2013).

Figure 1. Maximum Fraction of Balls in the “Heaviest” Urn: The Limiting Mean and Simulated Data ($n = 100$; 10,000 samples)Figure 2. Maximum Fraction of Balls in the “Heaviest” Urn: The Limiting Mean and Simulated Data ($n = 250$; 10,000 samples)Figure 3. Maximum Fraction of Balls in the “Heaviest” Urn: The Limiting Mean and Simulated Data ($n = 500$; 10,000 samples)

Figure 4. Maximum Fraction of Balls in the "Heaviest" Urn: The Limiting Mean and Simulated Data ($n = 1,000$; 10,000 samples)Figure 5. Maximum Fraction of Balls in the "Heaviest" Urn: The Limiting Mean and Simulated Data ($n = 2,000$; 10,000 samples)Figure 6. Maximum Fraction of Balls in the "Heaviest" Urn: The Limiting Mean and Simulated Data ($n = 5,000$; 10,000 samples)

Figure 7. Maximum Fraction of Balls in the "Heaviest" Urn: The Limiting Mean and Simulated Data ($n = 10,000$; 10,000 samples)Figure 8. Maximum Fraction of Balls in the "Heaviest" Urn: The Limiting Mean and Simulated Data ($n = 20,000$; 10,000 samples)